HARVARD DATA SCIENCE REVIEW
Just Accepted
DOI: 10.1162/99608f92.ad8ebbd4
ISSN: 2644-2353

Toward a Theory of AI Errors

Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI

Prof. Veronica Barassi
School of Humanities and Social Sciences, University of St. Gallen, Switzerland

Abstract

The rise of generative AI confronts us with new and key questions about AI failure, and how we make sense of and learn how to coexist with it. Whilst computer scientists understand AI failure as something that we can learn from and predict (Amodei et al., 2016; Yampolskiy 2018, p. 142), in this article I argue that we need to understand AI failure as a complex social reality that is defined by the interconnection between our data, technological design, and structural inequalities (Benjamin, 2019; Broussard, 2023) by processes of commodification (Appadurai & Alexander, 2019) and by everyday political and social conflicts (Aradau & Blanke, 2021). Yet I also show that to make sense of the complexity of AI failure we need a theory of AI errors. Bringing philosophical approaches to error theory together with anthropological perspectives, I argue that a theory of error is essential because it sheds light on the fact that the failures in our systems derive from processes of erroneous knowledge production, from mischaracterisations and flawed cognitive relations. A theory of AI errors, therefore, ultimately confronts us with the question about what types of cognitive relations and judgements define our AI systems, and sheds light on their deep-seeded limitations when it comes to making sense of our social worlds and human life.

Keywords: AI failure, AI hallucination, theory of errors, foundational models, anthropology

Just Accepted

1. Introduction

In September 2022 the American cognitive scientist Douglas Hofstadter, together with his colleague David Bander prompted GPT-3 with amusing probes such as: "What's the world record for walking across the English Channel?" *or* "How many parts will a violin break into if a jellybean is dropped on it?". In no time, the AI tool generated the most surreal, hilarious, non-factual responses: "The world record for walking across the English Channel is 18 hours and 33 minutes" and "A violin will break into four parts if a jellybean is dropped on it". In an article that appeared on *The Economist*, Hofstadter defines these responses as "cluelessly clueless", meaning that GPT-3 "had no idea that it had no idea", and added that "the system has no sense, it's random babbling" (Hofstadter, 2022, para. 5). Since then, these non-factual, surreal, and hilarious responses of generative AI tools have proliferated. The day that Google launched Bard in 2023, for example, the factual error made by the model in relation to the James Webb Space Telescope went viral and was followed by expert debates, media commentaries and TV appearances aimed at discussing the accuracy of generative AI models and whether we could trust them.

What is becoming clear is that our new technologies are not only astonishing and incredibly powerful, but they are also fallible, inaccurate, and at times completely nonsensical. To make sense of this fallacy, the term 'AI hallucination' has become widespread in the tech industry, the media, and the public imaginary. Google, Microsoft and OpenAI all publicly engaged with the problem. An internal Microsoft document stated that: "these systems are built to be persuasive, not truthful […] This means that outputs can look very realistic but include statements that aren't true." (Weise & Metz, 2023). The so called 'hallucination problem' is – as I am writing this article – one of the key problems of generative AI, which as Sundar Pichai, the CEO of Google and Alphabet has said "no one in the field seems to have solved" (Klein, 2023).

Yet we need to ask ourselves: are these AIs really hallucinating? Why are we using the word hallucination in the first place? This is the question that the critical scholar Naomi Klein asks herself in an article that appeared on *The Guardian* in May 2023. She argues that the concept of AI hallucination risks anthropomorphizing and giving agency to our generative AI tools and that, "by appropriating a word commonly used in psychology, psychedelics and various forms of mysticism, the tech-industry is feeding the myth that by building these large language models, we are in the process of birthing an animate intelligence" (Klein, 2023, para. 2). As Placani (2024) has shown, anthropomorphism in relation to AI can be particularly problematic for two main reasons. On the one hand, it tends to exaggerate and misrepresent AI capabilities by attributing human-like attributes onto systems that do not possess them. On the other hand, it distorts moral judgments about AI, such as those concerning AI's moral character and status as well as judgments of responsibility and trust in AI (Placani, 2024).

Although problematic, because of its anthropomorphizing dimension, the term AI hallucination is particularly interesting because it talks directly about Toffler's (1970) theory of *Future Shock*. As Toffler (1970) argued, the continuous and accelerating changes brought about by technological developments leave us: "progressively disoriented and increasingly incompetent to deal rationally with our new environments (2022 [1970], p. 11)". We rely on buzzwords; on the cultural definitions offered to us by the tech industry or the media and we do not have the time to critically reflect on their meaning or implications. The concept of AI hallucination is a vivid example of this.

In this article I want to take that time back. I want to engage with current debates about hallucinating systems and catastrophic AI failures, by considering their social and historical dimension, beyond buzz words. Technological failures, errors, and glitches have long existed and have often been key definers of technological development and innovation - from Enrico Fermi's misreading of nuclear fission to accidents in rocket science (Galison, 2005). Yet - as different scholars have pointed out - debates about failure have often been marginal in the study of technological developments (Ballatore & Natale, 2023; Magaudda & Balbi, 2024; Schiffer, 2019), and this is particularly true with reference to AI technologies (Aradau & Blanke, 2021).

Just Accepted

This article will argue that the rise of generative AI – and especially the rise of foundation models - confronts us with new and critical questions about AI failure, and its role in society. This is not only because – as we shall see - the greater homogenization of our systems can expose these models to greater degrees of failure (Bommasani et al., 2021), but also because the rapid integration of these systems in different sectors of everyday life can lead to potentially catastrophic consequences for individuals, organizations, and society. In reflecting on this transformation, the article will argue that AI failures cannot be understood as accidents that we can fix or learn from, but they need to be understood as complex social realities that are defined by the economic, social, and political relationships embedded within and around our technologies. Understanding AI failures as complex social realities, presupposes that we take a step further and critically reflect on the relationship between AI failure and *error as an epistemological concept*. By bringing the philosophical theory of error together with anthropological theories about the cultural specificity of knowledge, the article aims to question some of the limits of our models and open up the debate for development of a critical theory of AI errors.

2. Too Big Not to Fail: The Emergence of Foundation Models, Homogeneity, and the Problem of AI Failure

The study of technological failures, although largely neglected has been a fascinating field of enquiry at the intersection of different disciplines. For example, whilst the archaeologist Schiffer (2019) focused on the history of technological flops, and investigated the multiple technologies that were created but failed to launch (e.g. the electric car) communication scholars focused on the cultural narratives of failure in computer software (Ballatore & Natale, 2023; Magaudda & Balbi, 2024). In the human computer interaction literature, instead, we find discussions about technological failure as the inability of technological design and innovation to account for the situatedness of human social behavior. These discussions go a long way, from Suchman's seminal work (1987) of almost 40 years ago to the discussion in the literature on human-centred design. In sum, the research at the intersection of different disciplines clearly shows that technological failure, can and should be understood as a complex issue that covers a wide variety of technologies and social implications.

Although technological failure can be understood in broad terms, this article focuses on AI failures, where AI is understood here as involving multiple systems and applications that rely on machine learning, deep learning, natural language processing, text analytics, voice recognition, speech recognition, and/or computer vision. Only recently social scientists have turned their attention to social implications of AI technologies that fail (Ananny, 2022, 2023, 2024; Aradau & Blanke, 2021; Barassi & Patra, 2022; Munk et al., 2022; Rettberg, 2022). Within this literature, the focus has been on AI systems employed for automated decision making and predictive analytics, and the discussion on foundation models and generative AI is still relatively new (Ananny, 2024).

Of course, the issue of AI failure is not new at all. The AI safety expert Yampolskiy (2018, p. 42), for instance, has argued that if we look back into history, we find different example of 'intelligent agents' that have failed, from an advice software that deduced inconsistent sentences using logical programming 1958 to a nuclear attack early warning system that falsely claimed that an attack was taking place in 1983. In the AI Incident Database (McGregor, 2021) we find more examples of intelligent systems that – over the last twenty years – have failed: from the patriot missile system that led to a missile being launched by the American Airforce against a UK Tornado fighter jet in 2004 to the Amazon's cataloguing system that censored LGBT+ books. Within the AI safety community, some studied these AI failures as 'accidents' that emerge from poor design of real-world AI systems (Amodei et al., 2016), others have instead looked at the 'malicious use' of AI such as in the case of AI weapons or deepfakes (Brundage et al., 2018; Chesney & Citron, 2019; Scharre, 2016).

Whilst AI failure has long existed, the rise of generative AI is bringing something new to the table. This is because as Bommasani et al. (2021) have argued our generative AI technologies rely on foundation models that are more exposed to failure than others (2021, p. 117). Bommasani et al. (2021)

Just Accepted

coined the term 'foundation model' to refer to a family of models, which include also Large Language Models (LLMs) but are not limited to 'language', and which are aiding the overall development of generative AI (2021, p. 48). Foundation models rely on transfer learning and the ability to take the "knowledge" learned from one task (e.g., object recognition in images) and apply it to another task (e.g., activity recognition in videos). They also rely on self-supervised learning, which is far more scalable than supervised learning (2021, pp. 3-5). This implies not only that these models are trained with billions of parameters, but they can also be adapted to carry out a wide range of tasks despite not being trained for them.

According to Bommasani et al. (2021) before 2019, self-supervised learning with language models was essentially a subarea in NLP, and progressed in parallel to other developments in the field. After 2019, self-supervised learning with language models became more of a substrate of NLP, as BERT has become the norm. In the last few years, thus, we have witnessed not only a technological transformation and a change in AI methods, but also a social and cultural transformation. In fact, as Bommasani et al. (2021) argue, it was at this time that AI developers have started to accept and push forward the idea that a single model could be used for a wide-variety of tasks, and could be actually adapted to different contexts and situations (2021, p. 5).

The rise of foundation models, as the scholars explain, has led to an unprecedented level of homogenization. Today most NLP rely on just one or two foundation models such as BERT, RoBERTa, Bard, T5, etc. The most interesting example is of course the Google search, which now depends on foundation models like BERT as one of its signals (2021, p. 7). Whilst homogenization can be of fundamental importance because it ensures the power, scale and scope of our AI technologies, it also exposes our systems to a greater degree of failure. This is because foundation models rely on self-supervised learning and transfer knowledge, and hence all flaws and biases in the model can be inherited by all adapted models (2021, pp. 5-6). It is for this reason that Bommasani et al. (2021) argue that the threat of failures is particularly pertinent for foundation models in contrast to other kinds of AI and that as we integrate these models in our society, these failures can easily become catastrophic, if we think about failures in terms of warfare systems (resulting in unwanted discharge of weapons, possibly igniting a conflict) or critical infrastructure (accidental destruction of critical energy or agricultural capabilities) (2021, p. 117).

The rise of generative AI thus confronts us with new and critical questions about AI failure, and how we make sense of it. Whilst computer scientists have been mostly preoccupied with preventing failure and finding solutions (Amodei et al., 2016; Yampolskiy, 2018, p. 142), I believe that we need to take a step back and critically reflect on the social and cultural implications of AI failures, and on what they tell us about the technologies and the society we are building.

3. AI Failure as Complex Social Reality: From Commodification to Everyday Experiences and Conflicts

Our worlds are surrounded by failure. Failure is habitual, ubiquitous, it's a part of our daily lives, of what it means to be human. We should not be surprised thus if our technologies and AI systems fail, as they are the products of who we are and the reflection of the vulnerability of our material worlds. Yet, as Appadurai and Alexander (2019) have argued, this habitual, every day, unremarkable, and common aspect of failure has been largely denied. Rather than being understood for its habitual and anthropological dimension, failure has often been sensationalized and romanticized. From Zuckerberg's preaching about 'the freedom to fail' to the self-help manuals of the business world, failure is understood in a romantic way as a positive force that drives self-improvement and progress (Appadurai & Alexander, 2019, p. 27). This same approach to failure can be found in the AI safety literature mentioned above, which often seeks solutions and professes the importance of learning from past events of failures to build more robust, more ethical systems (Amodei et al., 2016; Yampolskiy, 2018, p. 142).

Just Accepted

As critical data and AI scholars have shown, unfortunately it is not that simple. AI failures are not accidents or glitches that can be fixed or prevented but they are the product of a long history of racism, sexism, ableism, and ageism in our data and technological designs (see Benjamin, 2019; Broussard, 2023; Noble, 2018; Wachter-Boettcher, 2017). Over the last few years, researchers have shown how these failures have been impacting human rights and amplifying inequality (Amoore, 2020; Barocas & Selbst, 2016; Crawford, 2021; Eubanks, 2018). They have also shown how difficult it is to shed light and challenge these failures, because these systems are often black boxes, non-transparent and unaccountable, (O'Neil, 2016; McQuillan, 2016; Pasquale, 2016). AI failures thus cannot be understood as single events or accidents that we can fix or learn from, but they need to be understood as complex social realities that are defined by the economic, social, and political relationships embedded in our technologies. AI failures need thus to be understood as sociotechnical-constructs and as *relations* between humans and machines (Ananny, 2022, p.4).

Understanding AI failure as complex social realities also presupposes that we consider how the tech industry commodifies failure. Most of our tech-products are launched with the idea that they will eventually fail, that all services will need to be updated and that new models will have to be launched (Appadurai & Alexander, 2019, p. 30). It is not only the big-tech that commodifies failure. As Bialski (2024) shows with her ethnographic work *MiddleTech: Software Work and the Culture of Good Enough*, software is too often built to be just "good enough" to function, and workers create and sustain a complex culture of 'good enoughness'. With the rise of generative AI, we are actually seeing the most bizarre articulations of how this culture of 'good enough' clashes with greater questions about the ethical, social, environmental and economic impacts of the models we are building. Only few months after launching ChatGPT, Open AI CEO Sam Altman admitted in an ABC interview that "the model will confidently state things as if they were facts that are entirely made up." (ABC, March 16, 2023), and few months later he testified in the US Congress about the need for more regulations. The release of Google Bard can also be seen as reflective of the business idea of the Silicon Valley that it is better to release a not-perfected technology rather than lagging the AI race (Jones, 2023).

Understanding AI Failure as complex social reality hence presupposes that we shed light on the fact that AI failures lead to a multiplicity of conflicting beliefs, emotions, fears, anxieties, practices, discourses, policies and solutions in our society. One of the most fundamental questions of our times thus is to understand the multiple ways in which different sections of society are responding, negotiating and finding solutions to AI failures. It is for this reason that in 2020 I have decided to focus on the social conflicts and debates that were emerging as a result of the use of AI technologies for predictive analytics and automated decision making and I have launched *The Human Error Project: AI, Human Rights and the Conflicts over Algorithmic Profiling*. Within the project I used the term 'the human error of AI' as an umbrella concept to shed light on different aspects of AI failures when it comes to human profiling (bias, inaccuracy, unaccountability) (Barassi, 2020). The idea of the project was rooted in the understanding that when we think about machine learning and error there is much to gain if we develop a 'history of the present' approach. Such an approach can enable us to understand how AI error "becomes a problem, raises discussion and debates, incites new reactions, and induces a crisis in the previously silent behavior, habits, practices, and institutions." (Foucault, 2001 as cited in Aradau & Blanke, 2021, p. 2).

In our project we thus decided to focus on three areas of society which were actively involved in the production of the public discourse around AI failures and errors in Europe: the media, civil society organizations and tech entrepreneurs. For all these different sections of society, we gathered data primarily through three main methodologies: critical discourse analysis, organizational mapping, and the collection of 105 in-depth interviews. Whilst during our three-year research project we gathered much valuable and interesting data on how the media is sensationalising and framing discourses around AI errors (Barassi et al., 2022), on some of the challenges of civil society organisations in dealing with AI bias and, on the failed promises of AI ethics models amongst entrepreneurs, I do not wish to discuss the findings here. Rather I want to draw on the theoretical insights of the project. Here I want to focus

Just Accepted

on the relationship between AI failure and error, and critically reflect on why we need a theory of AI errors.

4. Why We Need a Theory of AI Errors for Generative AI

One interesting aspect of the emerging body of literature on AI failures, is that there is lack of a theoretical scrutiny over the relationship between failure and error, and the two terms are often loosely defined or used interchangeably. This is evident not only if we consider the AI safety literature where scholars do not engage with a definition of failure or error (Amodei et al., 2016; Brundage et al., 2018; Scharre, 2016; Yampolskiy, 2018) but also if we consider the works of critical scholars and social scientists. In the works of critical data scholars like Noble (2018), Benjamin (2019) and Broussard (2023), for instance, the scholars draw on STS approaches to failure to demonstrate that these are not technical glitches but the reflection of the structural inequalities and biases. Within their work, thus, the scholars do not engage in a theoretical discussion about failure. Similarly, in the works of the sociologists and anthropologists like Munk et al. (2022) and Rettberg (2022), who call for an empirical study of algorithmic failure, the authors do not discuss the anthropological understanding of failure (see Appadurai & Alexander, 2019) let alone with error. Within the works of Aradau and Blanke (2021) and Ananny (2023, 2024) failure and error are more systematically theorised. Yet also in these works, the scholars do not engage with the relationship between failure and error or with critical questions about the *theory of error* in philosophy.

In philosophy, error has been defined as an epistemological concept, because it relates directly to our perception of reality. The basic understanding is that error emerges not from the absence of knowledge but because as humans we perceive reality through moral judgments. For Kant, for example, error does not simply arise from the comprehension of objects in isolation or from the senses devoid of judgment but from the act of making judgments about objects[1] (Kazim 2017, p. 58). Errors occur when there's a 'mismatch' between our judgments and reality (Lopes, 2010, p. 214). There are many similarities between Kant's approach to error and Indian Philosophy. In the theory of knowledge of Indian Philosophy (Khyativāda) error is often treated in relation to concepts such as perception, illusion, or hallucination (Matilal, 1981). The theory of errors is present in most schools of Indian philosophy, yet during our research we found the perspective of the Nyāya School, which was founded in the third century BC, particularly fascinating (Bhattacharjee, 2021; Chatterjee, 2016; Dravid, 1996).

According to the Nyāya School, error is defined by mis-characterization, it is connected to the idea of "illusion" and refers to "the cognition of an object as other than what it really is" (Chatterjee, 2016, p. 33). For instance, if we confuse a shell for silver or a rope for a snake. In this framework the error/illusion is not given by the fact that the elements are not real, but by a perception that has been mediated by an erroneous cognitive relation (Dravid, 1996, p. 38). In this framework the Nyāya School positions that error: "is not the absence of knowledge, but an object is judged to be what is not", error then refers to "the elements of reality being wrongly combined" (Bhattacharjee, 2021, pp. 960-961). I am not a philosopher, and unfortunately cannot dwell in the complexity of the relationship between Kant and Asian Philosophy, but others have (Palmquist, 2010). Yet I find that the definition of error as *illusory inference* (Lopes, 2010, p. 214) created by the elements of reality being wrongly combined crucial to understand AI errors.

Of course, there is a clear difference between human and machine errors, especially with reference to 'moral judgement'. Yet it is undeniable that like humans, our systems trace connections and patterns through cognitive relations and *combine different elements of reality to reach a specific conclusion* or perspective. The philosophy of error teaches us that when our AI systems fail, they fail because – like it happens to humans – they combine the elements of reality in ways that are value ridden and not objective or accurate. This understanding is particularly crucial to think about so called

[1] In western philosophy Kant's theory of error has inspired different debates from debates in cognitive science about truth and objectivity (see Roberts, 2011) to debates about moral judgements, and whether moral judgements could ever be non-erroneous (Mackie, 1977), which are not discussed here.

Just Accepted

generative AI hallucinations, which are the product of a probabilistic and erroneous combination of elements of reality. Yet this understanding can and should be applied also to other AI systems, such as predictive analytics technologies that focus on patterns and regularities that are often arbitrary in the understanding of human experience and behavior (e.g. the case of the HR software that came to different conclusions about the personality of candidates on the basis of whether they had a picture hanging in the back or not, see Barassi et al., 2022).

A theory of AI errors, thus, ultimately confronts us with the question about the process of knowledge production in our AI technologies. What types of cognitive relations and judgements define our systems? How are these erroneous, valued-ridden or deluded? When we ask ourselves these questions we quickly realize the limits of our machines. AI systems are fantastic mathematical achievements, which see patterns in data that humans would not normally see. Yet we need to appreciate that their process of knowledge production – or as McQuillan would say their algorithmic seeing (McQuillan, 2016) is limited because they rely on mathematical and statistical reasoning (Broussard, 2023, p. 12). In debates about human vs machine intelligence this understanding is key because it shows that our AI tools – for how powerful they will become - will not be able to replicate human forms of knowledge production. As cognitive scientist Gigerenzer (2023) has shown, one key definer of mathematical and statistical reasoning in our machines is their ability to solve and assess predictable problems. Yet our worlds are deeply unpredictable and messy, and we need different forms of intelligence to make sense of them. Whilst the knowledge that AI technologies generates can be incredibly insightful it is limited to a specific form of reasoning, and hence cannot fully understand the cultural complexity of human experience and behaviors.

When we think about the relationship between AI failure and error, we should also understand that the problem is not only a cognitive one, but also anthropological one. Over the last decade, research on AI bias and ethics has shown that our systems do not grasp 'the pluriverse' of human experience (Escobar, 2018) and that in most cases these systems are trained with biased data which is defined by simplistic, and reductionist understandings of how humans act, behave and show emotions (Crawford, 2021). The real problem is that we simply do not have generalized data that can account for the cultural diversity and complexity of human life. The anthropological literature on human intelligence (see Boas, 1911), the self and personhood (Morris, 1994; Strathern, 2002), or on human values and actions (Graeber, 2002) has clearly shown that human intelligence, experience, and behavior is not only unmeasurable, but also context-specific and culturally complex. Still today – as anthropologists have shown - we have very little grasp of the incredible cultural diversity of human experience, let alone generalized data that can train our models.

To shed light on this point it is crucial that we look at example of language. As some scholars have shown only a very small number of the over 7000 languages of the world are represented in our language models (Bender et al., 2021; Bommasani et al., 2021; Joshi et al., 2020). Earlier research into this problem can be found in the debates about the methods used to train GPT-3, where according to Brown et al. (2020), 93% of the training data was in English. The reason for this is not only that these technologies have originated in English speaking worlds but also that there is a lack of resources in other languages. In fact, as different scholars have argued there is more raw text data available in English (Joshi et al., 2020) and - as Johnson et al. (2022) have shown - whilst English represents 5% of world languages, it constitutes 44.9% of languages represented on the internet. Of course, over the last years we have witnessed important developments in multilingual models, which try to compensate for the lack of resource data, and thus challenge the problem of language representation. A good example can be found in the work that is being done to ensure greater representation of African languages in NLP and initiatives such as AFRIComet (Wang et al., 2023). Yet the problem is far from being solved. According to Nicholas and Bhatia (2023) multilingual language models often rely on machine-translated text that can contain errors or terms not in use by native language speakers. In addition to this these models cannot work equally well in all languages and often fail to account for the contexts of local language speakers (2023, p. 10). This leads of course to all sorts of implications in terms of fairness and equal distribution (see especially Bender et al., 2021).

Just Accepted

The problem goes further than the mere question of representation and lack of diverse perspectives and talks directly about the complex relation between language and culture. Most of our systems – including multilingual models - are trained on textual data (being it written, visual or audio). Yet the question that we need to pose ourselves, and that brings us back to the discussion on error as epistemological concept, is whether we can really understand our languages, and cultural worlds through text. The anthropologist Franz Boas would disagree. In his work on *Race, Language and Culture* (Boas, 1982) he argued that language training through textual materials "gives one sided presentation of linguistic data" (1982, p. 200), not only because texts often represent structural inequalities but also because language is not only about 'text' but about everyday embodied, relational practice. Although old and contested the work of Franz Boas shows that language is always situational and in continuous transformation and that texts were not capable of including the "possibilities of linguistic growth" (Boas, 1982, pp. 200-201).

The issue of language thus sheds light on all the complexities of our worlds and cultures, on how ignorant we still are about the world we inhabit and its processes, and on the impossibility of our systems to grasp these complexities and gaps in knowledge, even if we feed them with trillions of texts.

5. Conclusion

The rise of generative AI thus confronts us with new and key questions about AI failure, and how we make sense of and learn how to coexist with it. Developing a theory of AI errors is of central importance because it can enable us to accept and understand the limits, the habitual failures, and the erroneous forms of knowledge implicit in generative AI. AI failure and errors can bring about catastrophic outcomes for individuals and society. We need to learn to recognize them, coexist with them and deconstruct the illusion that we can fix them. AI systems are human-made and they are fallible just as we are. They will always be shaped by our cultural biases, our structural inequalities, and our lack of knowledge in specific areas.

Anthropologists have long sought to address the fact that our interpretation of the world and our systems of knowledge production are inevitably biased and limited (Clifford & Marcus, 1986; Geertz, 1973). From an anthropological perspective, there is nothing we can really do to 'correct' or combat this. The only thing we can do is to commit to self-reflexivity, rich ethnographic data, and unpredictability. As Strathern (2002) has shown, one of the richness of anthropology lies in the fact that it highlights the unplanned, the unpredictable, and the disconnections (2002, p. 309). Knowing to deal with unpredictability, is of central importance because it enables us to appreciate that the generative AI revolution is embedded in a complex and messy network of social interactions, communication and negotiation between people, imaginaries, and technologies. As argued elsewhere (Barassi, 2015), I believe that it is from these interactions that are open-ended and complex that comes much of the creativity and energy of social life. AI failures thus - paradoxically – talk directly about this energy of social life. They are crucial to the techno-historical period that we are living, not only because they can enable us to open the public debate about AI (Annany, 2024), but also because they confront us with critical questions about the processes of knowledge production of our systems, their flawed judgements, and their limitations.

Just Accepted

Disclosure Statement

This article emerged in the context of a wider research project titled, *The Human Error Project: AI, Human Rights and the Conflict over Algorithmic Profiling*, led by Prof. Veronica Barassi and funded by the Basic Research Fund (GFF) of the University of St. Gallen.

References

ABC News. (2023, March 16). OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: 'A little bit scared of this'. ABC News. https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv. https://doi.org/10.48550/arXiv.1606.06565

Amoore, L. (2020). *Cloud ethics: algorithms and the attributes of ourselves and others*. Duke University Press.

Ananny, M. (2022). Seeing like an algorithmic error: What are algorithmic mistakes, why do they matter, how might they be public problem?. *Yale Journal of Law and Technology*, *24*(1), 342-364.

Ananny, M. (2023). Making Mistakes: Constructing Algorithmic Errors to Understand Sociotechnical Power. *Osiris*, *38*, 223–241. https://doi.org/10.1086/725146

Ananny, M. (2024). Making Generative Artificial Intelligence a Public Problem. Seeing Publics and Sociotechnical Problem-Making in Three Scenes of AI Failure. *Javnost - The Public*, *31*(1), 89–105. https://doi.org/10.1080/13183222.2024.2319000

Appadurai, A., & Alexander, N. (2019). *Failure*. Polity.

Aradau, C., & Blanke, T. (2021). Algorithmic Surveillance and the Political Life of Error. *Journal for the History of Knowledge*, *2*(1), 1–13. https://doi.org/https://doi.org/10.5334/jhk.42

Ballatore, A., & Natale, S. (2023). Technological Failures, Controversies and the Myth of AI. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 237–244). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00026

Barassi, V. (2015). *Activism on the Web: Everyday Struggles Against Digital Capitalism*. Routledge.

Barassi, V. (2020). *Child Data Citizen: How Tech Companies are Profiling Us from Before Birth*. The MIT Press.

Barassi, V., & Patra, R. (2022). AI Errors in Health? The problem of scientific bias and the limits of media debate in Europe. *Morals & Machines*, *2*(1), 34–43. https://doi.org/10.5771/2747-5174-2022-1-34

Barassi, V., Scharenberg, A., Poux-Berthe, M., Patra, R. & Di Salvo, P. (2022). AI Errors and the Profiling of Humans: Mapping the Debate in European News Media', Research Report (I) The Human Error Project: AI, Human Rights and the Conflict over Algorithmic Profiling, School of Humanities and Social Sciences and MCM Institute University of St. Gallen, St. Gallen, Switzerland. https://thehumanerrorproject.ch/ai-errors-mapping-debate-european-media-report/

Barocas, S., & Selbst, A. D. (2016). *Big Data's Disparate Impact*. Social Science Research Network. https://papers.ssrn.com/abstract=2477899

Just Accepted

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Bhattacharjee, G. (2021). Theory of Error and Nyāya Philosophy: A Conceptual Analysis. *International Journal of Research and Analytical Reviews*, *8*(3), 958–964.

Bialski, P. (2024). *Middle Tech: Software Work and the Culture of Good Enough*. Princeton University Press.

Boas, F. (1911). *The Mind of Primitive Man*. Palgrave Macmillan.

Boas, F. (1982). *Race, Language and Culture.* University of Chicago Press.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the Opportunities and Risks of Foundation Models (No. arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Broussard, M. (2023). *More than a glitch: confronting race, gender, and ability bias in tech*. The MIT Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Éigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. arXiv. https://doi.org/10.48550/arXiv.1802.07228

Chatterjee, S. (2016). *The Nyaya Theory of Knowledge: A Critical Study of Some Problems of Logic and Metaphysics*. Motilal Banarsidass.

Chesney, B., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(6), 1753–1820. https://heinonline.org/HOL/P?h=hein.journals/calr107&i=1789

Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing Culture: The Poetics and Politics of Ethnography* (Second Edition, 25th Anniversary edition). University of California Press.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Dravid, N. S. (1996). The Nyāya-Vaiśeṣika Explanation of Illusion. *Journal of Indian Philosophy*, *24*(1), 37–48. https://www.jstor.org/stable/23447914

Accepted

Escobar, A. (2018). *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. Duke University Press.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Foucault, M. (2001). *Fearless Speech*. Semiotext(e).

Galison, P. (2005). Author of Error. *Social Research: An International Quarterly*, *72*(1), 63–76. https://doi.org/10.1353/sor.2005.0032

Geertz, C. (1973). *The Interpretation of Cultures. Selected Essays*. Basic Books.

Gigerenzer, G. (2023). *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*. Penguin.

Graeber, D. (2002). *Towards an Anthropological Theory of Value: The False Coin of Our Own Dreams*. Palgrave Macmillan.

Hofstadter, D. (2022, September 2). Artificial neural networks today are not conscious, according to Douglas Hofstadter. *The Economist*. https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). *The Ghost in the Machine has an American accent: value conflict in GPT-3*. https://doi.org/10.48550/ARXIV.2203.07785

Jones, R. (2023). Google Released Its Bard AI Before It Was Ready—But That Might Be Better Than Waiting. *Observer*, March 3rd. https://observer.com/2023/03/google-released-its-bard-ai-before-it-was-ready-but-that-might-be-better-than-waiting/

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Kazim, E. (2017). *Kant on Conscience: A Unified Approach to Moral Self-Consciousness*. BRILL.

Klein, N. (2023, May 8). AI machines aren't 'hallucinating'. But their makers are. *The Guardian*. https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein

Lopes, C. (2010). Truth, Falsehood and Dialectical Illusion: Kant's Imagination. In S. R. Palmquist (Ed.), *Cultivating Personhood: Kant and Asian Philosophy* (pp. 2015-216). de Gruyter.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Penguin.

Just Accepted

Magaudda, P., & Balbi, G. (2024). Theorizing Failure in Digital Media. Four Eclectic Theses. *Annals of the International Communication Association*, 1–14. https://doi.org/10.1080/23808985.2024.2326056

Matilal, B. (1981). Error and Truth: Classical Indian Theories. *Philosophy East and West*, *31*(2), 215–224. https://doi.org/10.2307/1399141

McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

McQuillan, D. (2016). Algorithmic paranoia and the convivial alternative. *Big Data & Society*, 3(2), 1–12. https://doi.org/10.1177/2053951716671340

Morris, B. (1994). *Anthropology of the self: the individual in cultural perspective*. Pluto Press.

Munk, A. K., Olesen, A. G., & Jacomy, M. (2022). The Thick Machine: Anthropological AI between explanation and explication. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517211069891

Nicholas, G., & Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis* (No. arXiv:2306.07377). arXiv. https://doi.org/10.48550/arXiv.2306.07377

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. https://www.degruyter.com/document/doi/10.18574/nyu/9781479833641.001.0001/html

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Palmquist, S. R. (2010). *Cultivating Personhood: Kant and Asian Philosophy*. de Gruyter.

Pasquale, F. (2016). *The Black Box Society: The Secret Algorithms That Control Money and Information* (Reprint edition). Harvard University Press.

Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00419-4

Rettberg, J. W. (2022). Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis. *Big Data & Society*, *9*(2), 20539517221131290. https://doi.org/10.1177/20539517221131290

Roberts, J. C. (2011). *The necessity of errors*. Verso Books.

Scharre, P. (2016). Autonomous Weapons and Operational Risk. Center for a New American Security. https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk

Schiffer, M. B. (2019). *Spectacular Flops: Game-Changing Technologies that Failed*. Eliot Werner Publications, Inc.

Just Accepted

Strathern, M. (2002). Abstraction and Decontextualization: An Anthropological Comment. In S. Woolgar (Ed.), *Virtual Society?: Technology, Cyberbole, Reality* (pp. 302–314). Oxford University Press.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

Toffler, A. (1970). *Future Shock*. Random House.

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. W. W. Norton & Company.

Wang, J., Adelani, D. I., Agrawal, S., Masiak, M., Rei, R., Briakou, E., Carpuat, M., He, X., Bourhim, S., Bukula, A., Mohamed, M., Olatoye, T., Adewumi, T., Mokayed, H., Mwase, C., Kimotho, W., Yuehgoh, F., Aremu, A., Ojo, J., … Stenetorp, P. (2024). *AfriMTE and AfriCOMET: Enhancing COMET to Embrace Under-resourced African Languages* (No. arXiv:2311.09828). arXiv. https://doi.org/10.48550/arXiv.2311.09828

Weise, K., & Metz, C. (2023, May 1). When A.I. Chatbots Hallucinate. *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Yampolskiy, R. V. (2018). Predicting future AI failures from historic examples. *Foresight*, *21*(1), 138–152. https://doi.org/10.1108/FS-04-2018-0034

Just Accepted

©2024 Veronica Barassi. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.